http://doi.org/10.17576/apjitm-2022-1101-05
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 11 No. 1, June 2022: 54 - 64
e-ISSN: 2289-2192

Received: 24 August 2021
Accepted: 29 November 2021
Published: 30 June 2022

A CLUSTER ANALYSIS OF POPULATION BASED CANCER REGISTRY IN BRUNEI DARUSSALAM: AN EXPLORATORY STUDY

DAPHNE TECK CHING LAI
OWAIS A. MALIK

ABSTRACT

Machine learning techniques have been mostly applied in gene expression cancer data. Socio-demographic data available in cancer registries could be explored, to get further insight into relationships between cancer types and their contributing factors. Moreover, less attention has been paid to analyse the mixed demographic data (numeric and categorical) from cancer registries and its association to the cancer types. The aim of this study is to identify subgroups of patients, having similar demographics characteristics, from the population based cancer registry in Brunei Darussalam and examine the prevalent cancer types in these subgroups. Four clustering algorithms are explored in the cluster analysis of Brunei Darussalam Cancer Registry; Two-step, Partitional Around Medoid, Agglomerative Hierarchical and Model-based. Gower distance was used for measuring similarity for mixed data types. To evaluate the clusters found; cluster distribution and Silhouette index were used for cluster quality, Cohen's Kappa Index for cluster stability and Cramer's V Coefficient for clinical relevance of clusters. Six distinct demographic subgroups were consistently found by three algorithms while model-based clustering solution were not considered for deeper analysis as highly imbalanced clusters were produced. The subgroups found have good quality clusters, moderate association with cancer types and high stability. The top three prevalent cancers associated with these subgroups were consistently identified using the three algorithms. Upon comparing the subgroups' ages during diagnosis, we identify possible screening behaviours of specific subgroups, suggesting for early screening awareness programmes. This study demonstrates the use of cluster analysis in a cancer registry to identify demographic subgroups that could suggest potential areas to develop targeted and improved healthcare management strategies.

Key-words: Population-based Cancer Registries, Cluster Analysis, Clustering Algorithms, Machine Learning, Two-Step, Partition Around Medoids, Agglomerative, Hierarchical, Model-based Clustering, Gower Distance

INTRODUCTION

Cancer is the leading cause of death, with 19.3 million new cases and 10.0 million deaths worldwide in 2020 (Sung et. al., 2021). Cancer accounts for approximately 13% of worldwide morbidities and mortalities with a 70% expected increase in the next two decades (Forman et. al., 2014). The worldwide cancer cases totalled up to 17.5 million and death cases of up to 8.7 million in 2015, with cases increased by 33% between 2005 and 2015 (Fitzmaurice et. al., 2017). Early detection and diagnosis of cancer is of great importance to facilitate appropriate clinical management and help increase survival rates. To support early detection and diagnosis, Machine Learning (ML) have been employed for tumor classification and cancer patients' prognosis, identifying critical features from complex datasets and predicting patient's survival time with good accuracy (Kourou et. al., 2015; Yu et. al., 2016; Hu et. al., 2021). Further, a better cancer control plan can be designed using ML as a decision support tool.

One common ML technique is the clustering of cancer data (e.g. gene expressions) for exploratory analysis (de Souto et. al., 2008). Clustering similar cases is useful for identifying underlying meaningful patterns in a dataset (Rokach and Maimon, 2005; Newcomer et. al., 2010; Khalil et. al., 2021). The similarities between cases are measured by their attributes values. These algorithms are mainly categorised into hierarchical, partitioning and model-based (Rokach and Maimon, 2005). The choice of a suitable algorithm and similarity/dissimilarity metric depends on the nature of the dataset. Clustering have been largely applied on cancer datasets for the discovery of cancer subtypes, breast cancer diagnosis and lung cancer stage identification (de Souto et. al., 2008; Garibaldi et. al., 2010; Dubey et. al., 2016; Demir and Karci, 2015; Ahmad, 2016; Kageyama et. al., 2021). A large-scale analysis of seven clustering approaches showed finite mixture of Gaussians and k-means clustering techniques for identifying the true structure in different cancer gene expression datasets (de Souto et. al., 2008). A consensus-based clustering methodology was applied to identify breast cancer subgroups using protein biomarkers (Garibaldi et. al., 2010). Recently, new algorithms (e.g. Firefly with golden ratio, foggy k-means) were used to improve clustering performance of cancer data (Demir and Karci, 2015; Yadav et. al., 2013). For clustering categorical data in breast cancer, a probabilistic distance measure was adopted (Ahmad, 2016). However, most studies used biomarker or gene expression cancer data for clustering analysis. To get insight into the relationships between cancer types and other factors, socio-demographic data available in cancer registries could be explored. ML have been useful in identifying cancer mortality patterns, and in predicting cancer survival and prevalence of other NCDs using the population’s socio-demographic characteristics (Malo et. al., 2007, Khalil et. al., 2021). Moreover, less attention has been paid to analyse mixed demographic data (numeric and categorical) from cancer registries and its relevance to cancer diagnosis (Malo et. al., 2007).

The use of socio-demographic variables in screening programs is already recognised as an area for good medical practice. ML cluster analysis provides for the ability to further refine such screening and diagnostic tools and algorithms based on socio-demographic variables and can also include geography. It allows for the establishment of such algorithms for less common cancers. Reliability and positive predictive value of such algorithms are improved as this can be based on real time data.

The aim of this study is to identify subgroups of patients, having similar demographics characteristics, from population-based Brunei Darussalam Cancer Registry (BDCR) and examine the prevalent cancers in these subgroups. Brunei Darussalam has an estimated population of 411,900 (2014 estimate) with racial groups Malays (65.8%), Chinese (10.2%) and others (24.0%). Her 2019 HDI is 0.838 - in the very high human development category - ranked 47 out of 189 (United Nations Development Programme, 2020). In Brunei, cancer is the leading cause of death, responsible for about 19% of total mortalities (Ministry of Health, Brunei, 2017). Brunei has relatively higher breast, lung, cervical and cancer rates as compared to other ASEAN countries (Kimman et. al., 2012); Lee et. al., 2012). In this study, we explored four clustering techniques; hierarchical, partitioning around medoids (PAM), model-based and two-step method with various similarity metrics and used a few cluster validity indices to evaluate cluster quality. The clusters found are evaluated based on stability and meaningfulness. Although cancer incidence in Brunei has been studied, the application of ML on BDCR has not previously been conducted. By identifying subgroups of cancer patients based on their demographic profiles, we hope to determine if there are unique cancer types prevalent in these subgroups. In doing so, these applications can provide supporting evidence to clinicians that may suggest potential improvement areas in healthcare management.

MATERIALS AND METHODS

FIGURE 1 shows the steps performed for clustering analysis of the cancer registry dataset.

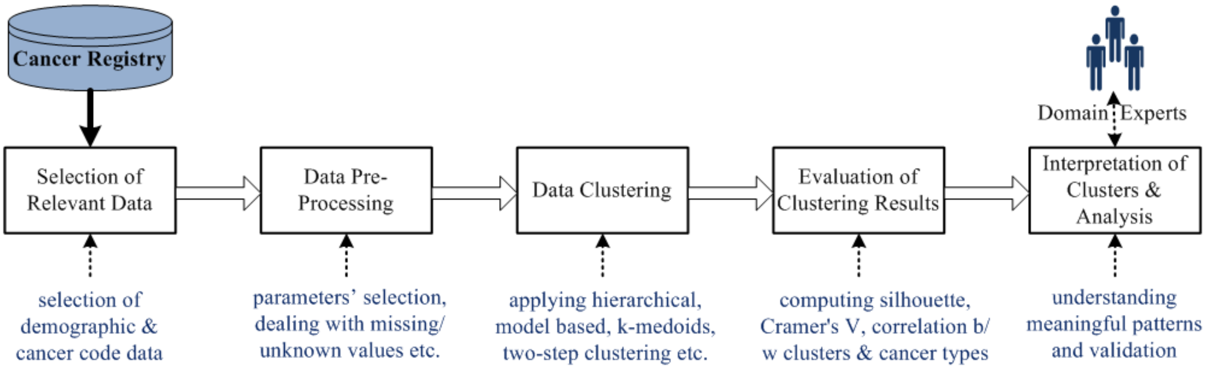
Cancer Registry
Selection of Relevant Data
Data Pre-Processing
Data Clustering
Evaluation of Clustering Results
Interpretation of Clusters & Analysis
Domain Experts
selection of demographic & cancer code data
parameters' selection, dealing with missing/ unknown values etc.
applying hierarchical, model based, k-medoids, two-step clustering etc.
computing silhouette, Cramer's V, correlation b/ w clusters & cancer types
understanding meaningful patterns and validation

FIGURE 1. Steps for Clustering Analysis of Cancer Registry Dataset

DATASET AND DATA PRE-PROCESSING

The BDCR contained 6327 patients' records (citizens and permanent residents) in Brunei with cancer incidence from year 2002 to 2016, after pre-processing. There were 5 parameters (4 demographic and 1 clinical) stored for each patient in the cancer registry. The clinical parameter is the ICD-10 code, International Codes of Disease type of cancer diagnosed.

Records having one or more missing/unknown values or outliers were removed. The frequency distributions of values for each nominal parameter were computed and examined to identify any discrepancy or errors in the data, which will be corrected manually or removed. Table 1 shows the demographics data used and the characteristics of then study cohort. The top 10 cancers in Brunei, in order of incidence are Breast (C50), Colorectum (C18-21), Lung, bronchus & trachea (C33-34), Cervix (C53), Non-Hodgkin lymphoma (C82-85, 96), Prostate (C61), Skin (C44), Ovary (C56), Stomach (C16), Liver (C22). To see the cancer types and their respective ICD-10 codes used in this study and number of cases based on cancer types and district, please refer to supplementary materials.

TABLE 1. Demographic Parameters Selected for Pre-processing and study cohort description

Demographic Parameter	***Data Type***	***Values***	***No. of cases***
Age of diagnosis	Numeric	(in years)	6327
Ethnicity	Nominal	Malay Chinese Others	5101 (80.6%) 1048 (16.6) 178 (2.8%)
Location (District)	Nominal	Muara Tutong Belait Temburong	4187(66.2%) 906 (14.3%) 1045 (16.5%) 189 (3.0%)
Gender	Nominal	Male Female	2752 (43.5%) 3571 (56.5)

CLUSTERING

HIERARCHICAL CLUSTERING

The agglomerative hierarchical clustering (HC) identifies cases with high similarity. Initially, each case belongs to its own clusters. Similarity is represented using linkages and distance measure. The two most similar clusters are combined to create a new cluster, replacing the two clusters. This process is repeated until one cluster is obtained. A cutoff point based on the chosen number of clusters is selected to obtain the cluster assignments of patients (cluster solution). HC is available in the R *cluster* package using *agnes* function (Maechler et. al., 2012) with flexible-beta agglomeration (Belbin et. al., 1992). The flexible-beta approach provides more details in cluster agglomeration using Lance-Williams dissimilarity (Lance and Williams, 1967). HC with flexible-beta was effective in finding clinically useful groups in health administration data (Cornell et. al., 2009). Instead of using Jaccard's coefficient as dissimilarity measure on binary data, we used Gower's measure on mixed data. Various linkages for updating similarity matrix of merged clusters were explored; Single, Complete, Average and Ward.D2 (Murtagh and Legendre, 2014), specified in *hclust* function. HC was applied to identify obesity subgroups in health and nutritional status survey data.

TWOSTEP CLUSTERING

TwoStep clustering (TSC) algorithm, found in BIRCH [25], is useful for analysing large datasets with both continuous and categorical data types. Clustering is performed in two-stages. First, the records are scanned sequentially and based on the distance criterion (Euclidian or log-likelihood), the current record is either merged with the previously formed clusters or created as a new cluster, using a modified cluster feature (CF) tree data structure [25] that handle both continuous and categorical data [26]. In stage two, HC is performed to generate the desired number of clusters using the subgroups formed from stage one. The number of clusters can be manually or automatically assigned in SPSS. The latter method initially calculates Schwarz's Bayesian Information Criterion (BIC) or Akaike's Information Criterion (AIC) of a specified range of cluster numbers, out of which, the most suitable number is determined.

MODEL-BASED CLUSTERING WITH BIC EVALUATION

Model-based clustering (MC) finds clusters of specific Gaussian finite mixture models in terms of cluster shapes and volumes (Fraley and Raftery, 2002). Using Expected-Maximisation (EM) algorithm initialised by HC, the model parameters are estimated. These models are then evaluated using BIC where the chosen model has the most favourable BIC score. The *Mclust* function from *mclust* R package (Fraley et. al., 2012) is used. Model-based clustering algorithms have been applied to evaluate uterine endometrioid carcinoma grade (Kageyama et. al., 2021).

PARTITIONING AROUND MEDOIDS

The Partitioning Around Medoids (PAM, also known as k-medoids) (Kaufman and Rousseeuw, 1987) works like k-means to minimize distances of (data) points belonging to the same cluster. Instead of regarding cluster average as cluster centre, PAM chooses the nearest point (to the cluster average) as its cluster centre. With continuous and categorical input data, PAM with Gower distance is chosen to find good representatives (cluster centres) that retain the structure for such data types. PAM is available in R (R Development Core Team, 2016) *cluster* package (Maechler et. al., 2012).

GOWER DISTANCE MEASURE

Gower distance measure (Gower, 1967) represents similarity of categorical data in the form of category matching, while maintaining similarity representation of continuous data in terms of geometric distance. BDCR contains both continuous and categorical data. For this reason, the

Gower distance measure is most appropriate. Euclidean and Manhattan distances would inaccurately represent similarities between categorical data as geometric distances, thus excluded in this study. Gower's distance measure d_{ij} compares two cases *i* and *j*, and is defined as follows:

$$d_{ij} = \frac{\sum_k w_{ijk} d_{ijk}}{\sum_k w_{ijk}}$$

where d_{ijk} denotes the similarity value of *k*th variable, and w_{ijk} is either 1 or 0 depending if differential variable weights are specified as *k*th variable's weight or 0 if the comparison is invalid. For categorical data, $d_{ijk} = 1$ if two cases *i* and *j* agree at *k*th variable and $d_{ijk} = 0$ if otherwise. For quantitative data,

$$d_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}$$

where r_k is the value range of *k*th variable. For continuous variables d_{ijk} ranges between 1, for identical values $x_{ik} = x_{jk}$, and 0, for the extreme values x_{max} - x_{min}.

For TSC, the Gower distance is an option in SPSS while for PAM and HC, the *daisy* method is available in R package, *cluster* (Maechler et. al., 2012). The *daisy* method takes data matrix as input and Gower measure is specified as parameter.

CLUSTER EVALUATION

The following techniques were used to evaluate the clustering solutions:

1. Cramer's V coefficient to measure association between cluster solution with cancer types;
2. Silhouette Index to measure cluster quality
3. Cluster distribution to determine cluster size balance
4. Cluster stability across different clustering algorithms (reproducibility) using Cohen's Kappa Index (κ)
5. Variability in cluster characteristics (distinctiveness) via cluster mode for nominal parameters

The Cramer's V coefficient (CV) is calculated using the clustering solution and the cancer types (based on World Health Organization (1992)'s ICD-10 codes) as inputs, both are nominal. A coefficient of near 0 indicates no or little association between the clusters found and type of cancer while a coefficient of near 1 indicates high association. The *assocstats* method in *vcd* package (Meyer et. al., 2020) is used to calculate CV, taking a contingency table as input, prepared using *xtabs* in R.

The Silhouette Index (SI) determines how well cases assigned to its clusters fit by comparing how similar the case is to cases in its own cluster and how similar to cases outside its clusters (Rousseeuw, 1987). An SI value range between -1 and 1 where SI value of near 1 indicate the data is well-clustered while near 0 indicate poor structure.

The *Kappa* method in *vcd* package calculates agreement levels between two clustering solutions, based on Cohen's Kappa Index (κ). A near one value indicates high agreement while near zero indicates otherwise. We use clusters' mode to study the specific demographic characteristics of the subgroups. With consideration of cluster association, distribution, quality and stability, cluster solutions containing subgroups with distinct (high variability) demographic characteristics are favoured.

RESULTS AND DISCUSSION

TABLE 1. Cramer's V association scores (CV) and respective p-value (likelihood ratio), Silhouette Index (SI), and cluster distribution of clustering solutions

Algo	***k***	***Cluster Evaluation***			***Cluster Distribution***				
		CV	***p***	***SI***	***1***	***2***	***3***	***4***	***5***
TSC	5	0.256	*	0.506	1488	766	1226	796	2051
TSC	4	0.291	*	0.501	1488	1562	1226	2051	-
TSC	3	0.124	*	0.235	3539	1562	1226	-	-
PAM	5	0.319	*	0.599	1974	2275	416	880	782
PAM	4	0.363	*	0.529	1974	2657	914	782	-
PAM	3	0.435	*	0.305	2756	2657	914	-	-
HC	5	0.318	*	0.632	1488	2051	1268	659	861
HC	4	0.361	*	0.583	1488	2051	1268	1520	-
HC	3	0.433	*	0.344	2756	2051	1520	-	-
MC	5	0.170	*	0.243	299	509	4867	240	412
MC	4	0.305	*	0.250	615	611	328	4773	-
MC	3	0.194	*	0.212	766	547	5014	-	-

* <0.0001

Table 2 shows the association score, SI and cluster distribution of clusters found by TSC, PAM, HC (with Ward.D2) and MC for 3 to 5 clusters. Table 3TABLE *2* shows the agreements between solutions from the algorithms for 4 and 5 clusters. Table 4 shows demographic characteristics of the 5 clusters found using TSC, PAM and HC.

CLUSTER EVALUATION USING CRAMER'S V AND SILHOUETTE INDEX

From Table 2, we observed that cluster solutions found using TSC, PAM and HC with 4 and 5 clusters have a CV value of above 0.25, including MC solution with 4 clusters. This indicates medium association (McHugh, 2012) between subgroups and cancer types. SI value of above 0.5 were obtained from TSC, PAM and HC solutions with 4 and 5 cluster, which strongly indicates a cluster structure exists in the data, whereas those with SI value of below 0.5 indicates poor hidden cluster structure. Based on the CV and SI values, solutions with 4 or 5 clusters are more favourable consistently in TSC, PAM and HC solutions.

CLUSTER DISTRIBUTION

On close examination of cluster distribution (see Table 2) and characteristics (see Table 4), TSC, PAM and HC clusters are found to be balanced and have distinct demographic characteristics. MC clusters are mostly dominated by one cluster with 4000 over patients, rendering them not useful for elucidating demographic profiles. Thus, we chose clusters found by TSC, PAM and HC for further analysis as they agree with other more and are balanced as compared to MC clusters.

CLUSTER STABILITY

Moderately high (substantial) agreement values (Landis and Koch, 1977) between clustering solutions are found with of above 0.6 (see Table 3) for TSC, PAM and HC, particularly for solutions with 5 clusters. This strongly indicates good cluster stability. MC solutions were found to have low agreement levels of below 0.3. Good cluster stability across solutions found by different clustering algorithms demonstrate confidence in the clusters found.

TABLE 2 Cohen's Kappa coefficient (κ) to measure agreement between TSC, PAM and HC solutions with 5 clusters and in brackets, 4 clusters.

	TSC		***PAM***		***HC***	
	Unweighted	***Weighted***	***Unweighted***	***Weighted***	***Unweighted***	***Weighted***
PAM	0.637	0.672	-	-	-	-
	(0.573)	(0.618)	-	-	-	-
HC	0.728	0.775	0.802	0.796	-	-
	(0.697)	(0.812)	(0.764)	(0.716)	-	-
MC	0.170	0.264	0.199	0.210	0.209	0.220
	(0.016)	(-0.002)	(0.095)	(0.095)	(0.074)	(0.074)

MC SOLUTIONS

While MC solution with 4 clusters produced CV of 0.305, the clusters' distributions are highly imbalance with the largest cluster containing 4773 patients while the smallest 328 patients. This was found in all MC clusters with 3 to 5 clusters. MC clusters have lower SI values when comparing with other algorithms. Furthermore, MC solutions were not reproducible by PAM, HC and TSC with κ of below 0.3, as observed on Table 3. For these reasons, MC clusters were not further analysed. The low values found CV, SI and κ measures indicate low association with cancer types, and low agreement with other solutions and these measures do not demonstrate high confidence in the found clusters, meaning the clusters found by MC were not supported by other clustering algorithms.

TABLE 3. Demographic Subgroup Found Through Cluster Analysis

Clus	***Consistent Demo Charac***	***Clus Algo***	***N***	***% of Total Cohort***	***Median Age (IQR)***	***Anchoring Demo Characteristics in Cluster(%)***	***Most Prevalent Cancers in Cluster(%)***
1	Muara Malay Female	TSC	2051	32.4	52(41,63)	Muara (100), Malay (100), Female (100)	Breast(24.3), Cervix(11.2), Colorectum(9.9)
		PAM	2275	36.0	51(41,62)	Muara(90.2), Malay(100), Female(100)	Breast(24.7), Cervix(11.0), Colorectum(9.5)
		HC	2051	32.4	52(41,63)	Muara (100), Malay (100), Female (100)	Breast (24.3), Cervix (11.2), Colorectum (9.9)
2	Muara Malay Male	TSC	1488	23.5	61(48,72)	Muara (100), Malay (100), Male (100)	Colorectum (17.5), Lung (15.1), Prostate (10.3)
		PAM	1974	31.2	60(47,71)	Muara (90.4), Malay (84.0), Male (100)	Colorectum (17.6), Lung (14.4), Prostate (9.4)
		HC	1488	32.5	61(48,72)	Muara (100), Malay (100), Male (100)	Colorectum (17.5), Lung (15.1), Prostate (10.3)
3	Muara Chinese Female	TSC	1226	19.4	60(50,73)	Muara (52.9), Chinese (85.5), Female (53.8)	Breast (16.5), Colorectum (15.7), Lung (11.7)
		PAM	416	6.6	55(47,69)	Muara (84.4), Chinese (89.9), Female (100)	Breast (33.2), Cervix (12.7), Colorectum (11.1)
		HC	659	10.4	56(48,69)	Muara (53.3), Chinese (86.8), Female (100)	Breast (30.5), Colorectum (12.3), Cervix (12.3)
	Belait	TSC	766	12.1	56(45,68)	Belait (80.7), Malay (100),	Breast(13.3), Lung(11.9),

4	Malay Female					Female (55.7)	Colorectum(11.7)
		PAM	880	13.9	58(49,70)	Belait(64.5), Malay(72.4), Female(100)	Breast (21.9), Colorectum(13.1) Lung (11.6)

5	Tutong Malay Female	TSC	796	12.6	58 (45,71)	Tutong (100),
						Malay (100),
						Female (54.5)
							Lung(13.9),
							Colorectum(12.3)
							Breast(11.9)

		HC	861	13.6	53 (43,65)	Tutong(50.4), Malay(100), Female(100)	Breast (22.5), Colorectum(10.8) Lung (9.6)

6	Belait Malay Male	PAM	782	12.4	67 (55,76)	Belait(61.0),
						Malay(67.9),
						Male(100)
							Lung(18.4),
							Colorectum(15.5)
							Prostate(10.0)

		HC	1268	20.0	64(51,74)	Belait (37.6), Malay (55.3), Male (100)	Colorectum(16.3), Lung(16.1), Prostate(8.7)

DISTINCTLY CHARACTERISED DEMOGRAPHIC PROFILES AND PREVALENT CANCERS

Table 4 shows the demographic profiles of clusters found using TSC, PAM and HC. By comparing the 5-cluster solutions, six distinct clusters are found with above 50% of cases in the cluster sharing common anchoring demographic characteristics. The most distinct clusters with more than 80% of its patients sharing common anchoring demographic characteristics are the Muara-Malay-Female and Muara-Malay-Male. For Muara-Malay-Female, it is the largest cluster with top three prevalent cancers breast, cervix and colorectum and median ages from 51 to 52. The second largest cluster is the Muara-Malay-Male, making up at least 84% of its cluster. The three prevalent cancers consistently found are colorectum, lung and prostate with median ages of 60 and 61.

In Muara-Chinese-Female, the prevalent cancers consistently found are breast and colorectum. While TSC identified lung as prevalent cancer in the group, PAM and HC identified cervix. This is due to the lower percentage of female in TSC cluster than in PAM and HC. The median ages found varies from 55 to 60. PAM and HC identified the sixth cluster with anchoring demographic characteristics Belait-Malay-Male.

The Belait-Malay-Female was identified by TSC and PAM, with prevalent cancers breast, colorectum and lung. The Tutong-Malay-Female found by TSC and HC identified prevalent cancers breast, colorectum and lung. The Tutong-Malay-Female identified by TSC and HC determined prevalent cancers breast, colorectum and lung.

Two algorithms found distinct clusters Belait-Malay-Female, Tutong Malay Female and Belait-Malay-Male consistently, creating an additional cluster to the original five clusters. These subgroups have consistent prevalent cancer types.

For Muara-Malay-Female and Muara-Chinese-Female, the prevalent cancers are breast, cervix and colorectum while lung was uniquely identified by TSC for the latter group. For Belait-Malay-Female and Tutong-Malay-Female, breast colorectum and lung were found to be prevalent cancers. Comparing age medians, the Muara-Malay-Female appears to be the youngest of the female groups (from 51 to 52) while other groups have medians from 53 to 60. For both Malay male groups (Muara-Malay-Male and Belait-Malay-Male), colorectum, lung and prostate were prevalent cancers. Muara-Malay-Male has lower median ages 60 and 61 while Belait-Malay-Male group has higher median ages 64 and 67. The differences in median ages of Malay Male and Female from the different districts may suggest Belait-Malay-Males and Belait-and-Tutong-Malay-Female are diagnosed at a later age, increase awareness for early screening in Belait and Tutong districts is to be strengthened. Further analysis is required to confirm this by studying the stages of cancer diagnosed among the groups.

Interestingly, Non-Hodgkin lymphoma was not identified as top 3 prevalent cancers in the subgroups, despite being ranked fifth while prostate cancer ranked sixth was found prevalent in male subgroups. This indicates that the Non-Hodgkin lymphoma cancer is less prevalent than the other top ranking cancers amongst patients with the identified demographic profiles.

Although use of descriptive statistics is possible to obtain similar results, all combinations of demographic (categorical) conditions will have to be permutated before performing the statistics on each characterised subgroups. This becomes particularly challenging to study with higher number of categories for each parameter, such as in nations with larger number of states/districts and ethnic groups. In Brunei, we will have 9 subgroups before considering the age parameter. Through the use of cluster analysis as initial exploratory multivariate data analysis tool, the distinct subgroups and their characteristics are quickly determined and the relationship between parameters can be deeply studied. By comparing the characteristics of each clusters in terms of the cancer prevalence, we can identify the districts where better screening programme as well as other cancer-related programme can be planned in, according to the needs of each district. This is particularly important in the nationwide hospital services and resource management.

LIMITATIONS

So far, we have provided an exploratory analysis to identify specific demographic subgroups and their top three most prevalent cancers, agreeable across at least two different clustering algorithms. The cluster analysis conducted did not take into account of the population statistics which may have contributed to cluster generation of more populous groups such as the Malays.

FUTURE WORK

Given that there are clusters found that may be not be agreeable among different clustering algorithms, consensus clustering is considered. A consensus clustering solution will be part of our future study. One to address the high majority of the Malay patient population, we intend to apply weight adjustments to the data before applying cluster analysis. In this work, we have used Gower distance which takes into account of both continuous and categorical data. As future work, we would like to investigate using metric learning techniques, comparing with other distance measures such as Jaccard and density-based clustering approaches.

CONCLUSION

In this preliminary work, cluster analysis is demonstrated to identify distinct demographic subgroups and their top 3 prevalent cancers. Breast and colorectum cancers were consistently found in the female subgroups, but, their median ages differ across subgroups with the Muara-Malay-Female having lower median ages than the other 3 female subgroups. Similarly for the male subgroup identified, the same top three prevalent cancers were identified with the Muara-Malay-Male having smaller median ages than Belait-Malay-Male. Interestingly, all cancers ranked 1 to 7 were identified as one of the top three prevalent cancers in the subgroups except ranked 6 Non-Hodgkin lymphoma, suggesting that it is less prevalent to patient with the identified distinct demographic profiles. The results from cluster analysis can be used to suggest potential improvement areas to current healthcare management strategies.

REFERENCES

Ahmad, A. (2016). Evaluation of modified categorical data fuzzy clustering algorithm on the wisconsin breast cancer dataset. *Scientifica*, *2016*.

Belbin, L., Faith, D. P., & Milligan, G. W. (1992). A comparison of two approaches to beta-flexible clustering. *Multivariate behavioral research*, *27*(3), 417-433.

Cornell, J. E., Pugh, J. A., Williams Jr, J. W., Kazis, L., Lee, A. F., Parchman, M. L., ... & Noël, P. H. (2008). Multimorbidity clusters: clustering binary data from multimorbidity clusters: clustering binary data from a large administrative medical database. *Applied multivariate research*, *12*(3), 163-182.

Demir, M., & Karci, A. (2015). Data clustering on breast cancer data using firefly algorithm with golden ratio method. *Advances in Electrical and Computer Engineering*, *15*(2), 75-84.

Dubey, A. K., Gupta, U., & Jain, S. (2016). Epidemiology of lung cancer and approaches for its prediction: a systematic review and analysis. *Chinese journal of cancer*, *35*(1), 1-13.

De Souto, M. C., Costa, I. G., De Araujo, D. S., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC bioinformatics*, *9*(1), 1-14.

Fitzmaurice, C., Allen, C., Barber, R. M., Barregard, L., Bhutta, Z. A., Brenner, H., ... & Satpathy, M. (2017). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA oncology*, *3*(4), 524-548.

Forman, D., Ferlay, J., Stewart, B. W., & Wild, C. P. (2014). The global and regional burden of cancer. *World cancer report*, *2014*, 16-53.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), 611-631.

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). *mclust v. 4 for R: normal mixture modeling for modelbased clustering, classification, and density estimation. Dept of Statistics, Univ*. of Washington Tech. Rep., 587.

Garibaldi, J. M., Soria, D., & Rasmani, K. A. (2010). Consensus clustering and fuzzy classification for breast cancer prognosis. In *ECMS* (pp. 15-22).

Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics*, 623-637.

Hu, X., Wang, Z., Wang, Q., Chen, K., Han, Q., Bai, S., ... & Chen, W. (2021). Molecular classification reveals the diverse genetic and prognostic features of gastric cancer: A multi-omics consensus ensemble clustering. *Biomedicine & Pharmacotherapy*, *144*, 112222.

Kageyama, S., Mori, N., Mugikura, S., Tokunaga, H., & Takase, K. (2021). Gaussian mixture model-based cluster analysis of apparent diffusion coefficient values: a novel approach to evaluate uterine endometrioid carcinoma grade. *European Radiology*, *31*(1), 55-64.

Kaufman, L., Rousseeuw, P. J., & Dodge, Y. (1987). Clustering by means of medoids in statistical data analysis based on the. *L1 Norm,~ orth-Holland, Amsterdam*.

Kimman, M., Norman, R., Jan, S., Kingston, D., & Woodward, M. (2012). The burden of cancer in member countries of the Association of Southeast Asian Nations (ASEAN). *Asian Pacific journal of cancer prevention*, *13*(2), 411-420.

Khalil, U., Malik, O. A., Lai, D. T. C., & King, O. (2021). Cluster analysis for identifying obesity subgroups in health and nutritional status survey data. *Asia-pacific journal of information technology and multimedia*, *10*(2), 146-169.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The computer journal*, *9*(4), 373-380.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Lee, M. Y., Telisinghe, P. U., & Ramasamy, R. (2012). Cervical cancer in Brunei Darussalam. *Singapore medical journal*, *53*(9), 604.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2012). Cluster: cluster analysis basics and extensions. *R package version*, *1*(2), 56.

Malo, E., Salas, R., Catalán, M., & López, P. (2007). A mixed data clustering algorithm to identify population patterns of cancer mortality in Hijuelas-Chile. In *Conference on Artificial Intelligence in Medicine in Europe* (pp. 190-194). Springer, Berlin, Heidelberg.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

Meyer, D., Zeileis, A., Hornik, K., Gerber, F., Friendly, M., & Meyer, M. D. (2020). Package ‘vcd’. *R package version*, 1-4.

Ministry of Health, Brunei. (2017). Health information booklet. http://www.moh.gov.bn/SitePages/Health%20Information%20Booklet.aspx

Murtagh, F., & Legendre, P. (2014). Ward’s hierarchical agglomerative clustering method: which algorithms implement Ward’s criterion?. *Journal of classification*, *31*(3), 274-295.

Newcomer, S. R., Steiner, J. F., & Bayliss, E. A. (2011). Identifying subgroups of complex patients with cluster analysis. *The American journal of managed care*, *17*(8), e324-32.

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological psychology*, *87*(1), 93-98.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53-65.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021).

Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249.

Team, R. C. (2013). R: A language and environment for statistical computing.

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352). Springer, Boston, MA.

United Nations Development Programme. (2020). Explanatory note on 2020 HDR composite indices Brunei Darussalam. http://hdr.undp.org/sites/default/files/Country-Profiles/BRN.pdf

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization.

Yadav, A. K., Tomar, D., & Agarwal, S. (2013, July). Clustering of lung cancer data using foggy k-means. In *2013 International Conference on Recent Trends in Information Technology (ICRTIT)* (pp. 13-18). IEEE.

Yu, K. H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., & Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, *7*(1), 1-10.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. *ACM sigmod record*, *25*(2), 103-114.

Daphne Teck Ching Lai
Owais A. Malik
School of Digital Science,
Universiti Brunei Darussalam.
daphne.lai@ubd.edu.bn, owais.malik@ubd.edu.bn